基于日志分析的民办高校大学生网络生活类型研究

陈润文　邱　勇　黄文彬　王　军
(北京大学信息管理系　北京 100871)

摘要:【目的】研究基于民办高校大学生的导航页面使用情况, 揭示民办高校大学生典型的网络生活类型及特点。【方法】通过对导航页面设置数据采集点, 获得民办高校大学生的点击行为和搜索行为数据, 进行统一建模和特征提取后, 利用聚类算法将其细分为几个有代表性的群体。【结果】将民办高校大学生划分为 6 个典型群体, 分别为重度视频型、昼学夜玩型、搜索追剧型、沉迷直播型、劳逸结合型、勤奋学习型; 民办高校大学生主要用网络看视频与直播, 仅有小部分学生利用互联网进行学习。【局限】PC 端导航页面行为仅能反映大学生网络生活的一部分, 且数据的时间跨度为两个月, 不能反映学生在期初和期末的行为差异。【结论】本研究实现了民办高校大学生群体中典型网络生活类型的识别, 这将有助于民办高校大学生特点和行为规律的发现和总结, 为提升高校服务管理水平提供参考建议。

关键词: 民办高校　日志挖掘　聚类分析
分类号: G35　TP311

1　引　言

近年来民办高校快速发展, 截至 2016 年底全国民办高校共计 742 所, 本专科在校生已达 616.20 万人[1]。与由政府拨款的公立高校相比, 民办高校得到的社会关注较少, 其学生在校生活状况不为民众所了解。对民办高校大学生的相关研究[2]表明, 他们的闲暇时间主要用于上网。本研究团队设计了大学生网址导航页面, 并与某网络服务商合作将其投放至湖北、云南、河北等地的民办高校, 基于用户点击和搜索数据对用户上网行为进行建模, 用 K-means 算法将用户聚类, 实现了民办高校大学生群体网络生活类型识别, 这将有助于民办高校大学生特点和行为规律的发现和总结, 为提升高校服务管理水平提供参考建议。

2　相关研究

2.1　民办高校大学生现状研究

目前, 对民办高校大学生群体的研究大多集中在其心理健康状况、就业择业状况、学习动机等领域, 而对该群体网上生活的相关研究则集中在上网动机和上网行为两方面。戚良燕等[2]对上海市 6 所民办高校大学生的闲暇生活进行问卷调查和访谈, 发现多数学生将闲暇时间用于上网, 但使用网络的主要目的是娱乐, 用于学习的时间较少。朱云汉[3]针对民办高校大学生网络学习情况, 利用问卷调查法了解学生的学习动机和学习行为, 发现仅有 27.5%的民办高校大学生认为学习是上网的主要动机, 其余人则对学习不感兴趣, 甚至否定大学教育的作用。此外, 民办高校大学生整体自觉性和自制力较差, 网上学习很快会转变为休闲娱乐行为。林红[4]对山东省 850 名民办和普通高校专科生的网络依赖状况进行问卷调查, 发现民办高校大学生的网络依赖程度显著高于普通高校大学生, 两类高校学生网络依赖均存在显著的性别和年级差异, 男性对网络的依赖显著高于女性, 二年级显著高于一年级。在上网动机方面, 64.57%的学生首选休闲娱乐, 且

通讯作者: 邱勇, ORCID: 0000-0002-2712-5204, E-mail: gallonqiuyong@pku.edu.cn。

经常因为休闲娱乐而导致其在时间管理上出现严重问题。问卷法和访谈法是研究民办高校大学生网络生活的主要方法，但它们的数据来自被试者的陈述，难以避免其主观性的影响。日志挖掘方法能较客观地反映其网络生活现状，但是由于数据较难获取，相关研究寥寥无几。

2.2 基于 Web 日志挖掘的网络生活类型研究

国内外利用Web日志挖掘方法分析网络生活类型的研究中，其分析流程与一般的数据挖掘类似，包括数据抽样、数据预处理、分析挖掘、解释评估。主要方法有统计分析方法、建模分析预测、序列模式发现[5]、关联规则挖掘[6]、聚类分析等。这些研究从用户网络生活的不同侧面反映用户类型及其特点。

在识别在线学习者类型方面，王敏[7]对国内某知名慕课平台的一门课程进行学习者行为日志挖掘，主要关注学习者自身属性特点和学习行为中的视频行为、课件习题行为、讨论区交互行为这几个方面所反映出的MOOC学习行为特征，利用相关分析探究其影响因素并用K-means聚类分析将学习者划分为4种类型，这助于学习者学习特点和行为规律的发现和总结，为教学设计的改进和学生的自适应学习提供指导。

在识别企业客户群体方面，张玉峰等[8]认为 Web日志挖掘为企业实现网络竞争情报动态分析提供了一种有效的途径，能为企业实施个性化服务提供依据；能发现潜在客户；能对企业网站进行优化以及实现企业客户群体分析和聚类。

在识别电商购物者需求类型方面，张文君等[9]通过浏览器日志挖掘探测消费者在电商平台下网购时的需求状态。该研究基于淘宝网女装页面类型访问序列对用户进行聚类分析，揭示出电商用户的需求状态，并帮助电商网站识别用户动态变化的需求状态。Prasad等[10]利用K-means和期望最大化两种聚类方法分析电商用户的注册数据和购物历史数据，为在线零售商店建立用户细分模型。Moe[11]则从在线商店的后台日志挖掘入手，通过分析用户访问页面的占比、用户访问的页面顺序等对用户的行为进行建模。

在识别网站访问者类型方面，于亚秀[12]提出一种综合用户浏览时间、点击次数的用户兴趣度度量方法，该方法基于用户的访问兴趣进行用户聚类，并在聚类的基础上构建针对用户兴趣的个性化网页推荐模型，这可以帮助网站管理者更好地了解网站访问情况，改善站点设计，提供自动化推荐，提高用户满意度。

总的来说，聚类分析是识别用户类型的主要方法，在选取用户聚类特征上先前的研究采用用户基本信息和行为信息相结合的方式，这为本文选取用户聚类特征提供了指导。在数据集的选取上，笔者则是基于自主设计的网址导航页面的点击和搜索数据，相比于之前的研究者在网上爬取的数据，可以有更多的操作空间，基于本研究的成果进一步迭代网址导航页面也能支持更深入的研究。在研究主题上，之前的研究聚焦于网络生活的某一个领域，如在线教育、网购消费等，能够在该侧面有较深入的洞察，但缺乏对网络生活全貌的分析，而本文通过分析大学生网址导航页面的行为数据来了解大学生的网络生活全貌。

3 研究设计

3.1 数据收集与导航网站设计

为了收集能反映民办高校大学生网络生活的日志数据，本文采取在其网络入口处嵌入导航页面并设置数据采集点记录用户日志的方法。导航网站的形式基本能够覆盖用户各方面的兴趣偏好，通过挖掘导航网站中产生的日志数据可以比较全面地反映该群体的网上生活状态。

网站链接和搜索功能是导航网站中最主要的两个功能，网址点击和搜索也是网络访问中重要的行为，因而本研究的导航网站设计包含搜索框与网站链接两个模块。其中，网站链接部分的网站类目体系需要覆盖民办高校大学生网络生活的各个方面。因此基于对现有的通用导航页面及大学生导航页面的调研，结合用户体验的角度确定其类目名称及各类下的网站数量，设计结果见4.1节。

之后，在该导航页面的搜索框及各个网址链接处设置数据采集点，以收集用户的网址点击日志与搜索日志，日志中包含用户ID、IP、时间、访问内容等字段。

3.2 日志记录的语义描述

学生的日志数据无法直接表示其网络生活特征，尤其是搜索记录中，搜索词的内容是不受控制的，因此需要对日志记录进行语义描述，将其表示为可以直接用于分析的形式。

本文将用户日志中的时间与访问内容进行抽象，

将用户行为日志统一表示为<UID, type, theme, hour>的四元组。其中, UID(用户 ID)用于标识用户, type 指该日志中用户进行了点击操作还是搜索操作, hour 指该日志发生的时间(小时), theme 指该日志中的访问内容(网站或搜索词)对应的语义类别。

在导航网站设计时已经给出网站类别, 但是这个类别是从用户体验角度考虑的, 类目的概念之间存在重叠的情况。而数据分析需要各类别下日志数量分布大致平衡, 因此需要对数据进行预分析, 将日志数量过多或过少的类别进行拆分或合并。

3.3　用户建模与聚类分析

结合日志数据的特征, 本文将用户上网偏好分为三个维度, 分别是客户端使用量、操作偏好和内容偏好, 结构如图 1 所示。其中, 客户端使用量能呈现大学生上网的依赖程度和时间规律, 操作偏好能呈现大学生上网的目的性强弱, 内容偏好则能刻画大学生上网的兴趣。

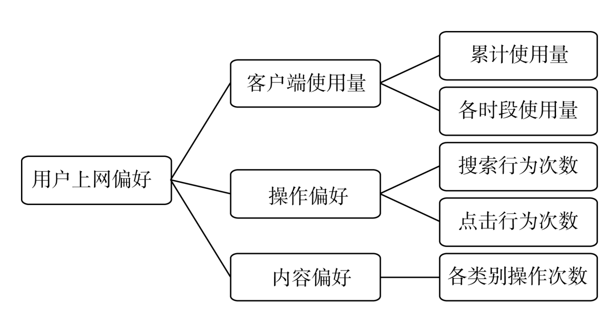


图 1　用户上网偏好模型

基于这些用户特征, 使用 K-means 聚类[13]即可得到民办高校大学生的典型类别。通过对各类别质心的特征数据进行分析, 可以得出其在客户端使用量、操作及内容偏好上的特征, 最终归纳出民办高校大学生网络生活的各个类型。

4　实验过程与结果

4.1　导航网站设计与投放

为了进行数据收集, 针对大学生群体设计了一个网址导航页面, 并与某网络服务商合作, 嵌入到该服务商在湖北、云南、河北等省份的民办高校网关客户端中。

基于对现有大学生导航网站的调研以及对民办高校大学生生活的基本了解, 笔者将大学生访问的高频网站按照其生活的维度, 从学、看、读、听、约、聊、玩、逛、挣等 9 个方面予以组织和呈现, 以全方位满足其上网需求。而在搜索框部分, 提供了百度、必应等多个搜索引擎选项。网址导航页面如图 2 所示。



图 2　大学生网址导航页面

4.2　数据获取与预处理

本文选取 2017 年 3 月 8 日至 2017 年 5 月 7 日作为分析时段。这段时间内, 用户共访问网站 6 657 181 次, 其中湖北、云南、河北三个省的用户访问次数分别占 39.17%、29.69%、17.64%。研究中通过预先在网页中设置数据采集点, 用百度统计平台收集到这两个月内用户在网站中的点击与搜索日志, 共得到 83 450 条原始记录, 日志数据结构如表 1 所示。

表 1　日志数据结构

点击		搜索	
字段	标签	字段	标签
downDate	日志日期	downDate	日志日期
time	时间	time	时间
UID	用户 ID	UID	用户 ID
URL	点击的网址	engine	搜索引擎
isHot	是否为热门	word	检索词
loginTime	登录时间	loginTime	登录时间
prov	省份	prov	省份
city	城市	city	城市

首先, 对这些数据进行初步的清洗, 删除用户 ID 缺失、时间字段出现偏误的记录, 并对重复的记录进行清理, 保留其中的第一条(同一用户在 5 秒内进行的完全相同的多次操作被认为是重复记录)。清洗后得到 67 762 条记录, 其中网站点击记录有 49 346 条, 搜索记录有 18 416 条。

将用户日志按时间字段进行整理，得到日志在一天内各个时段的分布情况，如图 3 所示。大部分日志产生在10:00-22:59之间，且12:00-12:59与16:00-21:59是用户日志产生的高峰，根据民办高校大学生作息时间分布的特性，可以看出在午间休息及下午课程结束之后他们会比较频繁地使用该导航网站。

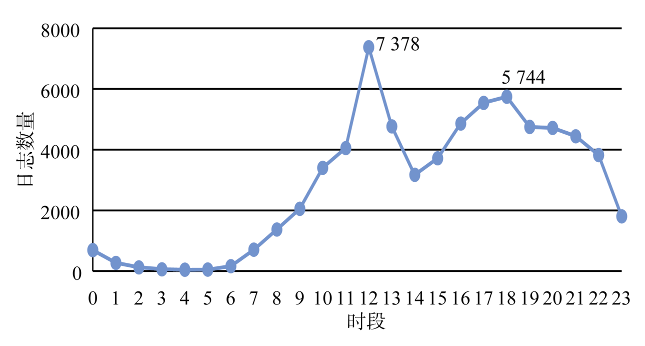


图 3　各小时用户日志数量

这些日志中包含 26 699 名用户，他们的日志数量分布如图 4 所示，对横纵坐标取对数之后可以发现用户日志数量符合幂率分布，拟合优度 R^2=0.9378。

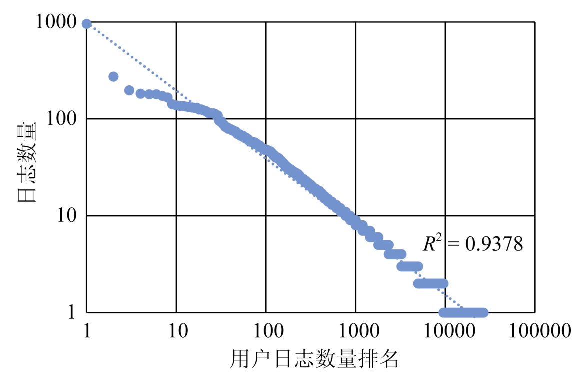


图 4　用户日志数量分布(对数坐标)

由图 4 可知，大量用户的行为日志数据较少，难以刻画用户网络生活的全貌，因此，选取日志数量不少于 20 条的用户进一步分析，筛选后共计 309 人，涵盖 14 814 条日志记录。

4.3　类目体系重构与人工标引

数据分析过程中，需要将导航页面的类目转化为适用于分析的明确的类别，然后据此对网站和搜索词进行标引。

(1) 部分类目的名称(如学、约、挣等)是从提升用户体验角度进行设计的，含义不够明确，需相应地调整为“学习”、“社交”和“工作”等。

(2) 原组织体系中概念过于宽泛的类别难以反映用户的具体行为。“看”这一类别下包含“看视频”、“看直播”等概念，将其拆分为“视频”类和“直播”类能够更细致地刻画用户在“看”这一类下的行为，而剩余的音乐、小说、漫画等则划为“娱乐”一类。

(3) 部分搜索词不属于导航页面组织体系中的任何一类，因此新增了“工具”、“学校”、“资讯”三类。如将“百度云”、“IE 浏览器”等工具型软件相关的搜索词划分为“工具”类；将“搜狐”、“网易”等门户网站与“李大钊”、“地铁一号线”等资讯查找相关的搜索词划分为“资讯”类；将“XX 大学信息门户”、“XXX 学院”等学校门户相关的搜索词划分为“学校”类。

新的类目体系共包含 11 个类别，其结构如图 5 所示，之后对搜索词和网站进行人工分类，每个网站和搜索词仅属于其中的一类。

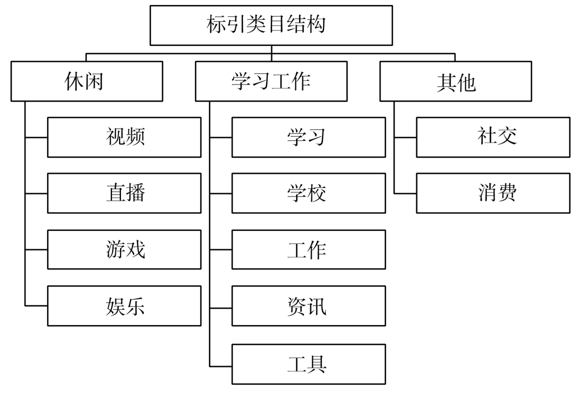


图 5　类目体系框架

4.4　网络行为模型构建

对搜索词和网站进行标引后，将用户行为统一表示为<UID, type, theme, hour>的四元组，具体标注样例如表 2 所示。

表 2　用户行为统一表示样例

UID	type	theme	hour
031101846031@campus	点击	消费	11
031101846031@campus	搜索	学习	11
031101846031@campus	搜索	学习	11
031101846031@campus	搜索	学习	13
031101846031@campus	点击	视频	12
031101846031@campus	点击	视频	13
031101846031@campus	点击	视频	13
031101846031@campus	点击	学习	16
……	……	……	……

按照图 1 所示的模型对用户特征进行提取。由于大学生的作息与课程安排密切相关，笔者对部分投放学校的日程进行调查，将时间分为上午(08:00- 11:59)、中午(12:00-12:59)、下午(13:00-17:59)、晚餐(18:00-18:59)、晚间(19:00-20:59)、夜间(21:00-次日 7:59)等 6 个时间段。

由于各用户的日志总条数差异较大，将操作类型、时间、内容类别的条数分别除以该用户的日志总条数，将其转化为各类别所占的比例，以统一数值大小，缩小日志总数差异造成的影响。之后，对日志总条数则取对数后标准化到[0,1]区间，也作为用户的一个特征。最终得到特征表样例如表 3 所示。

表 3　用户网络行为特征表样例

用户 ID	操作偏好		客户端使用量						
	点击	搜索	条数	上午	中午	下午	晚餐	晚间	夜间
031101846031	0.70	0.30	0.00	0.25	0.05	0.40	0.00	0.15	0.15
031102180309	1.00	0.00	0.09	0.29	0.14	0.14	0.29	0.14	0.00
31102195106	0.11	0.89	0.22	0.04	0.02	0.45	0.00	0.04	0.45
031102805624	0.85	0.15	0.00	0.10	0.15	0.30	0.10	0.15	0.20
31102814909	0.40	0.60	0.00	0.25	0.10	0.15	0.05	0.00	0.45

用户 ID	内容偏好										
	工具	工作	社交	视频	消费	学习	学校	游戏	娱乐	直播	资讯
031101846031	0.00	0.00	0.00	0.50	0.05	0.30	0.00	0.05	0.10	0.00	0.00
031102180309	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
31102195106	0.00	0.00	0.00	0.36	0.02	0.38	0.00	0.00	0.09	0.02	0.11
031102805624	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.05	0.80	0.00
31102814909	0.00	0.00	0.00	0.75	0.05	0.05	0.00	0.05	0.05	0.00	0.00

以 UID 为 031101846031 的用户为例，在操作偏好上，其所有行为中点击操作占比为 0.7，搜索占比为 0.3；在客户端使用量上，其条数为 0.00，指日志总条数在所有样本中最少，其中发生在上午、中午、下午、晚餐、晚间、夜间的操作分别占比 0.25、0.05、0.40、0.00、0.15、0.15；同样地，在内容偏好上的数值表示用户在该类别上的操作次数占比。

4.5　聚类结果及可视化

对用户进行特征提取后，采用聚类的方法将上网行为模式相似的用户进行归类，形成具有典型性的几类用户，从而更好地理解民办高校大学生中不同群体的上网行为和特点，为高校提升服务和管理能力提供支持。

基于上文中得到的用户特征表，使用 K-means 方法对用户进行聚类，分别取 4-7 类的结果进行比较，其中将用户聚为 6 类的效果最好。

图 6 显示了各类的人数及操作偏好，其中 6 个条形对应各类别，其总长度反映记录总数的多少，而左右的长度比代表点击记录数与搜索记录数的比例。可以看出，第 3、6 类对搜索有较强的偏好，而剩余 4 类则倾向于网站点击。

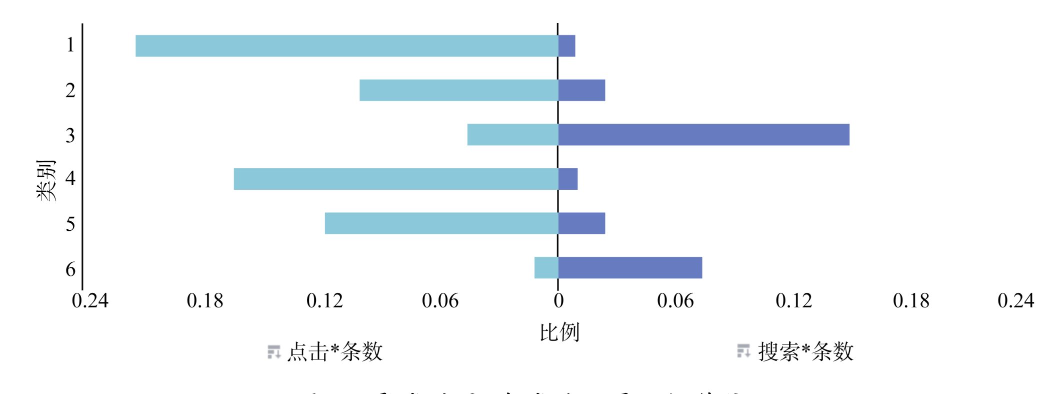


图 6　聚类结果–各类使用量及操作偏好

各类用户的操作在时间分布上的特性如图 7 所示。可以看出，第 3、5 类用户在上午的使用量相对较高，而第 2 类用户则在晚上 8 点之后有很高的使用量，各类总体上分布差异不大。

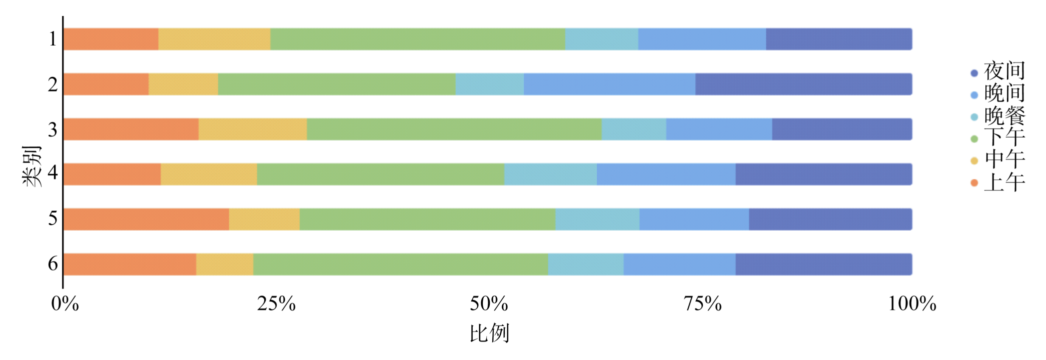


图 7　聚类结果—各时段使用比例

在内容偏好上各个类别则体现出较大差异，如图 8 所示。其中，第 1、2、3 类用户大部分操作与视频相关，而第 4 类对直播有很强的偏好，第 5、6 类用户则较少看视频、直播，而是更关注学习、工作、资讯等内容。

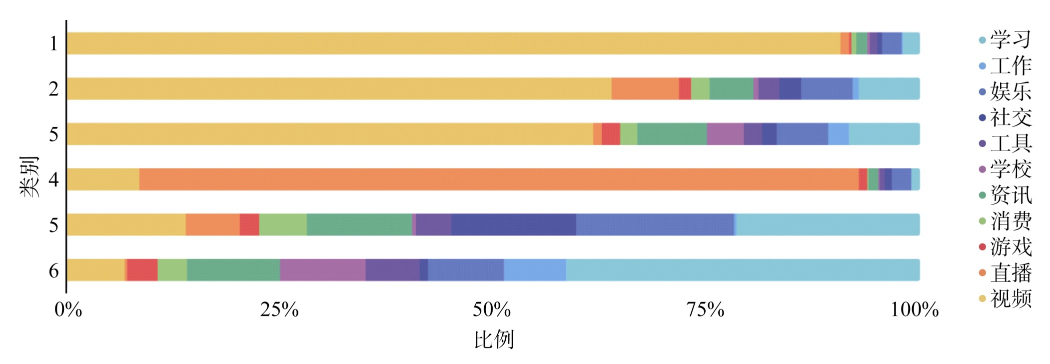


图 8　聚类结果—内容偏好

5　聚类结果分析

对 309 个民办高校大学生进行聚类后，得出 6 个具有比较明显特征的用户类别：重度视频型、昼学夜玩型、搜索追剧型、沉迷直播型、劳逸结合型、勤奋学习型。如图 9 所示，重度视频型占比最多，达到 26.9%；昼学夜玩型和沉迷直播型则各占到 20.7%。

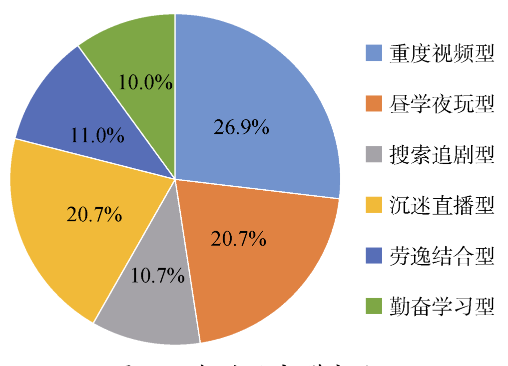


图 9　典型用户群占比

(1) 重度视频型

这一类用户有 83 人，数量在 6 类中最多，占所有用户的 26.9%。该类别最明显的特点就是对视频的偏好非常强，90.7%的日志记录为视频类别，其中最钟爱的网站是爱奇艺、腾讯视频、芒果 TV 等。并且，他们平均有 95.8%的操作是点击，仅 4.2%为搜索，体现出这类用户习惯于通过网页链接进入视频网站并在站内浏览或搜索，而非直接搜索剧集名称。此外，平均日志数量是各类中最高的，即很频繁地使用该导航跳转到视频网站，并观看影视剧与综艺节目。

(2) 昼学夜玩型

这一类用户有 64 人，占比为 20.7%。该类别同样对视频网站比较青睐，平均有 63.9%的日志属于视频类别，但其中除了爱奇艺、优酷等综合视频网站外，还有哔哩哔哩、AcFun 等弹幕视频网站，相对更关注动漫与游戏相关视频。除此之外，他们还对直播(7.8%)、

学习(7.0%)等内容有一定的偏好。这些用户另一个较为明显的特征是在晚上操作记录较多, 19:00 点之后的记录平均占一天内总记录的 45.6%, 而中午、下午的记录则明显比其他用户少。这说明他们的生活状态是白天以学习等事务为主, 而晚饭之后则相对空闲, 可以将大量时间用于娱乐。

(3) 搜索追剧型

这一类用户有 33 人, 占总人数的 10.7%。他们与"视频"相关的日志记录占比达到 60.7%, 即同样对视频内容有着较强的偏好, 而其他主题中则对"资讯"(8.1%)、"学习"(8.0%)关注最多。与上述两个类别不同, 这群用户的搜索日志占比平均达到 76.4%, 远超其网址点击的占比。分析该类用户的搜索词内容后可以发现, "人民的名义"、"武动乾坤"等电视剧名称及"向往的生活"等综艺节目名称有着较高的搜索次数。这说明该类用户虽然同样钟爱影视剧、综艺等内容, 但是有着更强的目的性。

(4) 沉迷直播型

这一类用户也相对较多, 占 20.7%, 共 64 人。他们对直播有着很强的偏好, 日志中平均有 84.4%的记录属于"直播", 而 "视频"仅占 8.6%。该类用户网址点击比例平均为 94.1%, 远高于搜索占比。这与直播网站的特性比较吻合, 通常直播平台的用户都是对平台甚至是特定主播有依赖, 因此会先进入相应直播网站, 而不是像影视剧等可以直接搜索剧集名称。此外, 这些用户在 18:00 点之后直播相关记录比例会上升, 可能是受主播的直播时间影响。而下午则有一些资讯、娱乐等内容, 这与大学生的作息时间有一定的联系。

(5) 劳逸结合型

该类有 34 名用户, 占总体的 11.0%。他们的内容偏好中, 占前 5 位的分别是学习(21.2%)、娱乐(18.4%)、社交(14.4%)、视频(13.9%)、资讯(12.1%)。他们有一定的娱乐需求, 但是也会将很多时间花在学习、资讯相关的网站上, 能够较好地平衡自己在学习与娱乐两方面所花的时间, 自制力较强。此外, 这一类用户在上午的记录是各类中最多的(平均占 19.6%)。

(6) 勤奋学习型

这一类用户数量最少, 为 31 名, 占 10.0%。他们对该网站的依赖性较低, 平均日志数量在各类中最少。在内容偏好方面, 排在第一位的是学习(40.1%), 而学校事务(10.0%)、工作(7.4%)也相对占比较高, 而与娱乐相关的各个类别则占比很少, 与其他类别差异较大。此外, 他们是搜索日志平均占比最高的一类, 达到 86.7%, 搜索次数较多的词有"智慧树"、"慕课网"、"365 大学网"等。这一类用户主要利用导航网站的搜索功能, 将其作为查找学习资料及其他资讯的便捷入口, 与其他类别呈现出较大的差异。

6 结 语

在上网内容方面, 大部分民办高校大学生的网上生活以娱乐为主, 将大量的课余时间用于看视频和直播, 呈现出重度的视频依赖, 这与之前民办高校大学生网上生活的研究结论相契合。即便如此, 仍然有近四分之一的民办高校大学生利用互联网进行学习。在上网时间方面, 民办高校大学生的上网时间段受学校的课程安排影响较大, 且大部分学生在零点之后无上网行为, 可见民办高校断电、断网的管理手段对其学生起到了约束作用, 这对于部分自制力较差的学生来说, 是一种有效减轻网络依赖的方式。综上, 笔者认为民办高校应该积极利用 MOOC 引导有重度视频依赖特征的大学生将兴趣转移到学习上, 这样既能提高学生的学习兴趣, 也能减轻民办高校的教育资源压力。

在以后的研究中将丰富和扩大日志数据的规模和维度, 更完整地刻画用户的上网行为; 在特征赋权上可以使用专家评估法, 从而得出更贴合现实的聚类结果, 提高研究的科学性和研究结论的普适性。

参考文献:

[1] 孙竞, 熊旭. 2016 年全国民办学校 17.1 万所 在校学生突破 4825 万人 [EB/OL]. [2017-01-18]. http://edu.people.com.cn/n1/2017/0118/c367001-29033819.html. (Sun Jing, Xiong Xu. The Number of Private Colleges has Reached 171,000 by 2016, with More than 48.25 Million Students [EB/OL]. [2017-01-18]. http://edu.people.com.cn/n1/2017/0118/c367001-29033819.html.)

[2] 戚良艳, 许月英. 上海民办高校学生闲暇生活调查与分析[J]. 浙江树人大学学报:人文社会科学版, 2010(4): 124-128. (Qi Liangyan, Xu Yueying. An Investigation and Analysis of Students' Leisure Life in Private Colleges and Universities in Shanghai [J]. Journal of Zhejiang Shuren University: Humanities and Social Sciences, 2010(4): 124-128.)

[3] 朱云汉. 论民办高校大学生网络学习行为[J]. 中国成人教育, 2015(14): 136-137. (Zhu Yunhan. Behavior of Private College Students' Online Learning [J]. China Adult Education, 2015(14): 136-137.)
[4] 林红. 民办与普通高校学生网络依赖状况的比较研究[J]. 青少年研究(山东省团校学报), 2008(6): 24-28. (Lin Hong. A Comparative Study on the Internet Dependence of Private College Students [J]. Youth and Adolescence Studies, 2008(6): 24-28.)
[5] 王继民, 彭波. 搜索引擎用户访问量模型[J]. 计算机工程与应用, 2004, 40(25): 9-11. (Wang Jimin, Peng Bo. Modeling Quantity of Users/Access for Search Engine [J]. Computer Engineering and Applications, 2004, 40(25): 9-11.)
[6] Srikant R, Agrawal R. Mining Quantitative Association Rules in Large Relational Tables [J]. ACM SIGMOD Record, 1996, 25(2): 1-12.
[7] 王敏. 基于行为日志数据的MOOC学习者学习行为分析研究[D]. 上海: 华东师范大学, 2016. (Wang Min. Research on MOOC Learning Behavior Based on Behavior Log Data [D]. Shanghai: East China Normal University, 2016.)
[8] 张玉峰, 何超. 基于 Web 日志挖掘的网络动态竞争情报分析研究[J]. 情报理论与实践, 2011, 34(9): 51-53. (Zhang Yufeng, He Chao. Research on Dynamic Competitive Intelligence Analysis Based on Web Log Mining [J]. Information Studies: Theory & Application, 2011, 34(9): 51-53.)
[9] 张文君, 王军, 徐山川. 电商用户需求状态的聚类分析——以淘宝网女装为例[J]. 现代图书情报技术, 2015 (3): 67-74. (Zhang Wenjun, Wang Jun, Xu Shanchuan. Clustering Analysis of Demand State of E-commerce Users - Taking Taobao Women's Clothing as an Example [J]. New Technology of Library and Information Service, 2015 (3): 67-74.)
[10] Prasad P, Malik L G. Generating Customer Profiles for Retail Stores Using Clustering Tech[J]. International Journal on Computer Science & Engineering, 2011, 3(6): 2506-2510.
[11] Moe W W. Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream[J]. Journal of Consumer Psychology, 2003, 13(1-2): 29-39.
[12] 于亚秀. 基于 Web 日志挖掘的个性化服务研究[D]. 上海: 华东师范大学, 2009. (Yu Yaxiu. Research on Personalized Service Based on Web Usage Mining [D]. Shanghai: East China Normal University, 2009.)
[13] Jain A K. Data Clustering: 50 Years Beyond K-means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666.

作者贡献声明:

黄文彬, 王军: 提出研究思路, 设计研究方案;
陈润文, 邱勇: 进行实验, 采集、清洗和分析数据, 论文起草;
陈润文, 邱勇, 王军: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。
[1] 陈润文, 邱勇. log.csv. 导航网站用户日志数据.
[2] 陈润文, 邱勇. tagging.xls. 搜索词、网站标引表.
[3] 陈润文, 邱勇. clustering.csv. 聚类结果表.

收稿日期: 2017-05-31
收修改稿日期: 2017-07-04

Analyzing Private College Students' Online Lifestyle with Web-logs

Chen Runwen　Qiu Yong　Huang Wenbin　Wang Jun
(Department of Information Management, Peking University, Beijing 100871, China)

Abstract: [**Objective**] This study reveals the private colleage students' typical online life styles based on their usage of a navigational Web portal. [**Methods**] First, we collected the click and search data of the navigation page specifically designed for students. Then, we modeled the data and applied the K-means cluster algorithm to categorize the student behaviors. [**Results**] We found six major behaviors among private college students. However, these students mainly use the Web to watch videos, while only a small number of students use the Web to learn. [**Limitations**] The size and dimensions of the data need to be expanded. [**Conclusions**] This study identifies typical online life styles of private college students, which could help schools improve their administraion and services.

Keywords: Private College　Log Analysis　Cluster Analysis